MiniGPT-4: Enhancing Vision-Language Understanding with Advanced Large Language Models

Year: 2023 | *Citations: 2623* | *Authors: Deyao Zhu, Jun Chen, Xiaoqian Shen*

Abstract

The recent GPT-4 has demonstrated extraordinary multi-modal abilities, such as directly generating websites from handwritten text and identifying humorous elements within images. These features are rarely observed in previous vision-language models. However, the technical details behind GPT-4 continue to remain undisclosed. We believe that the enhanced multi-modal generation capabilities of GPT-4 stem from the utilization of sophisticated large language models (LLM). To examine this phenomenon, we present MiniGPT-4, which aligns a frozen visual encoder with a frozen advanced LLM, Vicuna, using one projection layer. Our work, for the first time, uncovers that properly aligning the visual features with an advanced large language model can possess numerous advanced multi-modal abilities demonstrated by GPT-4, such as detailed image description generation and website creation from hand-drawn drafts. Furthermore, we also observe other emerging capabilities in MiniGPT-4, including writing stories and poems inspired by given images, teaching users how to cook based on food photos, and so on. In our experiment, we found that the model trained on short image caption pairs could produce unnatural language outputs (e.g., repetition and fragmentation). To address this problem, we curate a detailed image description dataset in the second stage to finetune the model, which consequently improves the model's generation reliability and overall usability. Our code, pre-trained model, and collected dataset are available at https://minigpt-4.github.io/.